DOCUMENT RESUME

ED 407 424 TM 026 446

AUTHOR Mathieu, Cindy K.

TITLE Basic Precepts in Test Construction: Recommendations from Various Measurement Textbooks.

PUB DATE 23 Jan 97

NOTE 24p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, January 23-25, 1997).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Achievement Tests; Difficulty Level; *Measurement Techniques; Selection; Standardized Tests; *Test Construction; Test Content; Test Items; Test Use

IDENTIFIERS *Paper and Pencil Tests

ABSTRACT

This paper presents six steps in test construction generally recommended by measurement textbook authors. The focus is primarily on paper-and-pencil achievement tests as used by class instructions, although the discussion touches on the construction of other types of assessment. The six steps are: (1) determine the test purpose; (2) determine the test content, including difficulty level; (3) determine the test format; (4) construct an initial pool of test items; (5) revise the items using input from qualified colleagues or other sources; and (6) pretest and revise the items. Developing a good test requires both effort and know-how. (Contains one table and eight references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Running head: PRECEPTS IN TEST CONSTRUCTION

ED 407 424

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
CINDY K. MATHIEU
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Basic Precepts in Test Construction:

Recommendations from Various Measurement Textbooks

Cindy K. Mathieu

Texas A&M University 77843-4225

Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX, January 23, 1997.

TM026446

Abstract

The present paper presents six steps of test construction generally recommended by measurement textbook authors: 1) determine the test purpose; 2) determine the test content; 3) determine the test format; 4) construct an initial pool of items; 5) revise the items; and 6) pretest and revise items. Particular attention is given to step four; with an emphasis on multiple choice, true/false, and essay item construction.

Basic Precepts in Test Construction:

Recommendations from Various Measurement Textbooks

A test has been defined as "a means of measuring the knowledge, skill, feeling, intelligence, or aptitude of an individual or group. Tests produce numerical scores which can be used to identify, classify, or otherwise evaluate test takers" (Gay, 1980, p. 111). As simple and straightforward as this definition sounds, some educators and other professionals may not fully appreciate the amount of effort and know-how that developing a good test requires. The purpose of the present paper is to discuss the recommended guidelines and basic steps in developing a good assessment instrument.

To begin, a test constructor must understand the solid foundation he/she will need on which to build a test. This foundation includes specification of purpose, standard conditions, consistency, validity, and practicality (Brown, 1983). First, all tests should have a clearly specified purpose. The specification of purpose will be discussed later in this paper, and it will be seen that determining what construct the test is designed to measure, how the results of testing will be used, and who will take the test, will all give direction to the test construction process (Brown, 1983). Maximizing standard conditions is a second goal in testing, as accuracy depends upon the control of errors in developing, administrating, and scoring the test. Minimizing irrelevant factors in testing is accomplished by making the test situation as similar as possible for all test takers. Third, consistency of the test scores is fundamental. Unless a test produces consistent scores, the scores will not mean much. Fourth, in order to

be interpretable, test scores must be valid. That is, they must represent the construct they were designed to measure, and nothing else. Finally, issues of practicality and efficiency should be built in to the test construction process. A test constructor should consider the time, money, and qualifications needed to administer, score, and interpret the test. The goal is to generally use the simplest procedures possible while maintaining the highest test quality (Brown, 1983).

With these fundamental concepts in mind, it should be noted that there are many different types of tests and many different ways to classify tests. For example, there are performance tests which may measure either maximal performance or typical performance. There are self-report methods, such as questionnaires, surveys, and interviews. Tests may be designed to measure achievement, personality, or aptitude. Tests may be given to individuals or to groups, and may be either speed tests or power tests. In addition, the criteria to which an individual's scores will be compared will be different if he/she is taking a norm-referenced or a criterion-referenced test. Because there are so many categories of tests, the remainder of the present paper will primarily focus on the most widely used tests, i.e., paper-and-pencil achievement tests as used by class instructors. A few key considerations in constructing other types of assessments will also be presented.

Steps of the Test Construction Process

Step One: Determine the Test Purpose

The first step in the test construction process is to determine the test purpose. A typical paper-and-pencil classroom test will be used to assess achievement, but other tests may be used for admissions, placement, or diagnostic decisions, or to measure personality or some other type of typical performance. A test may have more than one purpose, such as a test used to measure an individual's personality and also to help decide if the individual should be hired into a management position. Clarifying the purpose at the beginning of the construction process, then, will help ensure that the final test form will be useful for the purpose it is to serve (Crocker & Algina, 1986).

Step Two: Determine the Test Content

Second, the test constructor should determine the test content, or define the domain. In other words, the test purpose must be translated into operational terms. In achievement tests, it should be determined what information must be known by the test taker in order to pass the test. For example, a student in a test construction course may be required to know the steps of the test construction process in order to pass the final exam. In personality tests, the trait being measured must be defined as explicitly as possible, and the behaviors through which the trait will be manifested must be determined (Brown, 1976). For example, a test constructor wishing to develop an instrument to determine introversion/extroversion of a test taker must specifically describe the behaviors associated with being introverted and extroverted and develop the test to

measure those behaviors. In tests that are used to predict, such as the SAT or the ACT, the specific behaviors related to future success in college should be determined and tested.

A difficulty often encountered in this step of the process is selecting a small enough sample of items from a large domain to make the test an adequate yet practical assessment of the construct. For example, an achievement test on the ability to multiply decimals presents the problem of constructing all possible items from this domain. The possible number of items is obviously enormous, given the millions of number combinations available to the test constructor. Therefore, it is recommended that an item specification approach be used.

With this approach, items are written according to specifications so that they are interchangeable (Crocker & Algina, 1986). Using the example of a test on the ability to multiply decimals, some of the specifications given to write the stimulus may be: a) the problem should contain two decimal numbers written in horizontal form; b) the problem should be given with the instructions "Multiply the following numbers;" c) each of the numbers should have exactly two digits to the right of the decimal place; and d) each of the numbers should have one to three digits to the left of the decimal place. It should be evident that with the item specification approach, more than one individual can write the items and the items will still be equivalent in the knowledge they are measuring.

Another widely used system in developing test specifications is the use of a taxonomy of objectives. A taxonomy of objectives is a classification of objectives into a hierarchy of categories based on levels of complexity. The

classification is thought to be comprehensive and is ordered so that each category involves behavior more complex than the previous category and is a prerequisite to the next category (Gay, 1980). The two taxonomies that are used most often in constructing tests are based on the cognitive and the affective domains of behavior. These domains were established by Bloom, Engelhart, Furst, Hill, and Krathwohl in 1956.

The cognitive domain consists of six levels, or categories. Level one is Knowledge. This level involves test takers memorization, recall, and recognition of previously learned material. Test items from this category include the knowledge of facts, terminology, dates, and persons (Hopkins & Stanley, 1981). For example, a question asking the test taker to define the term validity is requires the test taker to exhibit skill at the knowledge level. Level two is Comprehension. At this level, a test taker is required to show understanding, not just memorization. For example, a comprehension level test item may ask the test taker to list all of the verbs in a paragraph. He/she would have to understand what a verb is, then, to answer this item. Level three is Application. A test item at this level would measure an individual's ability to use, or apply, an abstract concept to a specific situation. For example, an individual may apply his/her understanding of the general principles of inflation to the current U.S. economy. The fourth level is Analysis. This level includes the ability to break down a problem into its basic elements and to identify the relationships which exist between them. For example, a test taker may be asked to write a summary on a chapter entitled "Test Reliability." Level five is Synthesis, which involves

the ability to combine elements to form a unique whole, something that did not exist before. For example, an individual may be asked to add three lines to the first line of a poem, or to develop a plan to reduce unemployment in the U.S. Finally, level 6 is Evaluation. This level includes making judgments about the value of some idea. For example, a test taker may be asked to evaluate the current rating system of movies in the U.S.

The most important advantage of using the taxonomy of the cognitive domain when constructing tests is that it encourages the test constructor to develop items measuring higher mental processes. One of the most common defects in teacher-made tests which may result in very easy tests is the tendency to write too many items which measure only at the Knowledge level (Hopkins & Stanley, 1981). Although these items may be the easiest to construct, it is important to also have more difficult items in order to increase variability, and thus the reliability, of scores.

The objectives of the affective domain relate to feelings, attitudes, interests, and values (Bloom et al., 1956). This taxonomy consists of five levels. Level one is Receiving. At this level, an individual is sensitized to the phenomenon of interest and is paying attention to it. Listening attentively during a presentation of the test construction process is an example of this category. Level two is Responding. Individuals who are responding are making active responses, such as answering a question during a test construction presentation. Level three is Valuing. At this level, the individual acknowledges that the phenomenon has value. This would include continuing to think about

the test construction process after the presentation is over. The fourth category is Organization. An individual at this level incorporates the value into his/her value system in which some values are more important than others. For example, an individual may choose to spend more time studying the construction of achievement tests than the construction of typical performance tests. Finally, level five is Value Complex. At this level, an individual has completely internalized a value to the point that it causes him/her to behave in a consistent way. The affective taxonomy has not been used in education as much as the cognitive taxonomy has. However, the affective taxonomy may be useful to an individual wishing to construct interest inventories, questionnaires, or other self-report methods.

A table of specifications (or test plan) may be used at this point in the test construction process. The table is an outline specifying the coverage of the test. The table typically takes the form of a two-way grid with major content areas listed in one margin and cognitive processes on the other (Crocker & Algina, 1986). The purposes of the table are to determine how many and what sort of items need to be written and to determine at the end of the construction process if the final test form matches the test plan and therefore adequately samples the domain (Brown, 1983). For example, a test on descriptive statistics may have the following test plan, as determined by the test constructor:

	Knowledge	Comprehension	Application
Distributions	5%	10%	5%
Central tendency	15%	5%	5%
Variability	5%	5%	10%
Correlation	5%	5%	25%

Thus, this test constructor has decided to place the most emphasis on the test taker's ability to answer application items on correlations. Regardless of which areas the constructor wishes to emphasize, however, all percentages must add up to 100.

Step Three: Determine the Test Format

The third step in the test construction process is determining the test format. As noted earlier, most tests follow the paper-and-pencil-type format. However, this may not always be the best format for a test. A driving test, for example, must include a performance section designed to assess driving in order to adequately measure the driver's skill. Two considerations that arise in determining test format are the characteristics of the group to be tested and practical factors (Brown, 1983).

When taking into consideration the characteristics of the test takers, it should be determined how a test should be given to that particular group. For example, a test may need to be given orally if the test takers are young children or have limited reading skills. Practical factors involved with determining test format include the time needed to administer and score the test, the cost needed to develop, produce, and administer the test, and the qualifications needed by

individuals who will administer, score and interpret the test (Brown, 1983). For example, it is most practical to develop nationwide college admissions tests in a multiple choice format since scoring is much quicker and more efficient than most other test formats. It is also most practical to construct college admissions tests in such a way that large groups of individuals may take the exam at the same time. Although practicality is important, however, the overall quality of the test must always come first. In other words, if the choice is to use a more practical but poorer quality test or a less practical but higher quality test, the responsible test administrator will choose the latter.

Step Four: Construct an Initial Pool of Items

The fourth step in the test construction process is to construct an initial pool of test items. There is much common sense involved in constructing items, yet many of these basic guidelines are violated. For example, it is well known that well-constructed tests should present an entire item on a single page rather than continuing part of an item on a separate page, yet some test constructors violate this guideline. Therefore, a detailed discussion of this process should be helpful. The construction of multiple choice and true/false items will be discussed first, followed by essay item construction and inventory item construction.

A basic guideline for constructing multiple choice and true/false items is that the items should precisely measure the construct, not just "kind of" measure it. For example, if a test objective stated that a student should be able to compute split-half reliability, then an item only asking the student to list the

procedures involved in computing split-half reliability would not be appropriate. The only appropriate item in this case would involve having the student compute the reliability coefficient using a set of data (Gay, 1980).

Item novelty is an important concept in item construction. This guideline is sometimes violated by class instructors giving comprehensive examinations which include items written exactly as they were written on previous tests. In these and other situations in which a student is tested more than once over a concept, the test constructor should write new items each time. This helps ensure that the test taker's understanding of the concept is being tested, rather than his/her recall ability.

Item difficulty is another important consideration in writing norm-referenced multiple choice and true/false tests. In order for a test to produce high variability in scores, items must be neither too easy nor too hard. With this fact established, it should be understood that there are both correct and incorrect ways to make a test item more easy (or more difficult). An incorrect way to make an item easier is to write the response in a way that test takers could figure out the right answer even if they had not learned the material, as in the following example:

Who wrote Pride and Prejudice?

A. Elvis Presley

B. Jane Austen

C. Bill Clinton

D. Sitting Bull

The correct way to make an item easier is to decrease the difficulty of the required response, such as asking a question from a lower level of the cognitive domain taxonomy. On the other hand, the correct way to make an item more difficult is to write them in such a way that they require a deeper understanding of the concept (i.e., require a response from a higher level of the cognitive domain). The best way to develop plausible alternative responses is to construct them from common misconceptions or misinterpretations of the subject. For example, regarding the above item, a number of individuals like those to be tested may be asked who wrote Pride and Prejudice. Some of their wrong answers may be used as distracters for the item. The writing of plausible distracters, then, is fundamental in writing good multiple choice questions, and some experts would consider it to be the single most important skill in constructing these items (Hopkins & Stanley, 1981).

The following guidelines are also important when writing good multiple choice items: Do not provide clues to the correct answer. For example, use "a(n)" rather than "a" or "an" before the answer. The following item is an example of such an error:

An apparatus for measuring the work performed by a group of muscles is an

A. dynamometer.

B. ergometer.

C. spirometer.

D. tachometer (Gay, 1980, p. 245).

Even if an individual does not know what these instruments are, he/she will choose the correct answer (B) because it is the only one that begins with a vowel.

Avoid dependent items, such as the following:

If X + 8 = 10, then X = _____ and X + 1 = _____.

The correct answer to the first item is necessary in order to correctly answer the second item. Thus, if the first item is missed, the test taker will miss the second one even if he/she knows how to add.

Avoid irrelevant difficulty. Unless the test is meant to assess an individual's vocabulary, words like "walk" and "disagreement" should be used instead of "ambulate" and "altercation." Another example is the following item, which requires an individual to compute simple interest. It is unnecessarily difficult because of all of the decimal places: "If you borrowed $491.22 at 6.83% interest, how much interest would you pay for the first year?" (Gay, 1980, p. 238).

Avoid negatives, especially double negatives. The following item demonstrates the potential problem of using negatives:

Which of the following is not a requirement for a person desiring to be licensed as a real estate broker in the state of North Carolina?

A. An age of 21 years or older

B. No felony convictions

C. A score of 70% on the ETS Real Estate Brokers' Examination

D. U.S. citizenship (Gay, 1980, p. 239).

Although the answer is A, a test taker could easily become confused by the "not" in the question and the "no" in answer B. It is best to state the stem of the item in positive form, but if that is not possible and a negative must be used, it is recommended that it be emphasized in some way, such as by underlining or by using italics (Gronlund, 1977).

Do not test knowledge of trivial facts or direct quotes taken from a text. Such items will measure verbatim memorization rather than knowledge of the topic.

An item should have one and only one correct (or best) answer, unless otherwise specified. The following is an example of a faulty item:

The most serious disease in the United States is

A. cancer.

B. heart disease.

C. mental illness.

D. venereal disease (Gay, 1980, p. 240).

Obviously, the correct answer depends on how the test taker defines the word "serious." That is, does "serious" mean which disease causes more deaths,

which is the most prevalent, or which is most easily spread? The item should be reworded so that there is only one correct answer, such as asking which disease is the most prevalent in the U.S.

The stem should contain as much of the problem as possible. For example:

> An arachnid is
>
> A. an arthropod with 3 pairs of legs and antennae.
>
> B. an arthropod with 3 pairs of legs but no antennae.
>
> C. an arthropod with 4 pairs of legs and antennae.
>
> D. an arthropod with 4 pairs of legs but no antennae (Gay, 1980, p. 244).

This item is much less clear than:

> An arthropod with four pairs of legs but no antennae is called a(n):
>
> A. annelid.
>
> B. arachnid.
>
> C. insect.
>
> D. mollusk (Gay, 1980, p. 244).

The first item may unnecessarily confuse the test taker. This guideline may also be helpful in ensuring that the correct answer is not longer than the alternatives, which is a tendency of item writers and which provides a clue to the correct answer, especially to the test-wise student (Gronlund, 1977).

As noted above, the responses to the item should all be plausible. In addition, they should be grammatically consistent with the stem and the position

of the correct answer should be varied in a random manner (Gronlund, 1977). It is always better to have a few good distracters rather than a long list of poor ones. Responses can be arranged logically, by dates, or alphabetically. It is generally recommended that using responses such as "all of the above" and "none of the above" be avoided. "All of the above" can be detected as the correct answer even if the test taker only knows that any two of the alternatives are correct. It will be known that "all of the above" is not the correct answer if the test taker recognizes that at least one of the alternatives is incorrect. Test-wise examinees quickly catch on to this and their chances of guessing the correct answer from the remaining choices increases. Choosing a correct response of "none of the above" would indicate that a student knows which answers are wrong, but does not necessarily know what is right (Gronlund, 1977). The differential ability of test takers to attend to the item cues also tends to lessen the validity of test scores, since most classroom tests are intended to measure achievement rather than test-wiseness.

When writing true/false items, most of the guidelines for writing multiple choice items pertain, such as avoiding irrelevant difficulty, avoiding using most negative statements and all double negative statements, avoiding clues to the answer, avoiding making true statements longer than false ones, and having only one correct answer. In addition, words such as "always," "all," "never," and "none" should be avoided since they are associated with false items, and words such as "generally," and "usually" should be avoided since they are associated with true items (Gay, 1980). A disproportionate number of either true or false

statements should be avoided, and items should be arranged in a random order, rather than in some discernible pattern (such as T,F,T,F, etc.) (Gronlund, 1977).

The construction of high quality essay items requires the writer to be familiar with another set of rules or guidelines. It is important to review these guidelines, since many beginning class instructors erroneously believe that essay questions are the easiest to write. Actually, it is very difficult to construct clear, unambiguous essay questions. First, essay items are best used to measure complex learning objects, i.e., the synthesis and evaluation levels of the taxonomy of objectives (Gronlund, 1977). Second, the objective of the item must be as clear as possible. Words such as "compare and contrast," "evaluate," "analyze," "criticize," and "describe" should be used to clarify the task for the test taker. Writing the item in a less structured way will give the student more room to respond, and will make scoring much more difficult. The test-wise examinee, in particular, may claim to have "misunderstood" the intent of the question and answer according to what he/she knew (Hopkins & Stanley, 1981). Finally, it is not recommended that students be given a choice of questions on an essay test, such as choosing to answer any three of five items (Gay, 1980). It will be impossible to compare students' performances if they answer different items. Also, it decreases the instructor's ability to determine if the student has achieved the objectives. Since the student will choose to answer the questions he/she knows best, it remains unknown whether he/she learned the rest of the information.

When constructing inventory items (i.e., those requiring a Likert or semantic differential-type response), the following guidelines are some of those given by Crocker and Algina (1986):

1. Do not use statements that are factual or capable of being interpreted as factual.
2. Avoid statements that are likely to be endorsed by almost everyone or almost no one.
3. Try to have an almost equal number of statements expressing positive and negative feelings.
4. Avoid use of indefinite qualifiers such as *only, just, merely, many, few,* or *seldom.* (p. 80)

The last guideline can be demonstrated by the following example: "I have many friends," and "I make friends easily" may be poorly constructed items because of the many ways test takers may interpret the meaning of the statements. That is, individuals may have different interpretations of what constitutes a "friend," what number of friends would constitute "many" friends, and what making friends "easily" would mean. Therefore, care must be taken to develop items that are less ambiguous by clearly defining all terms.

Response set is a particular problem for those constructing self-report inventories. Response set is a tendency for test takers to respond in a fixed or stereotyped way when items consist of two or more possible response choices (Aiken, 1976). Achievement tests do not tend to be as susceptible to the problem of response sets, since the only ways a test taker can distort the scores

of these tests is to not respond, to deliberately answer wrong, or to guess. Two response sets that have been found to occur with self-report inventories are the acquiescence and the social desirability sets. The set to acquiesce involves the tendency to agree with a statement when the test taker has no informed basis for agreeing or disagreeing (Brown, 1976). For example, students may consistently rate instructors highly on evaluations, even if the course was not designed to show a certain skill of the instructor (such as giving a high rating on lecturing skills when the course was primarily based on guest speakers and student presentations). The social desirability response set means that items rated as being socially desirable are endorsed with greater frequency than items rated as being socially undesirable (Brown, 1976). For example, an item such as, "I have strange and peculiar thoughts" will be endorsed less frequently by test takers, especially those who understand the social undesirability of agreeing with such a statement. One technique used to minimize response sets is to use a forced-choice format. A forced-choice item consists of two or more statements that are equal in social desirability but are different in content and validity (Aiken, 1976). The following is an example of a forced-choice item:

Do you prefer your home to look

A. "lived in" and comfortable

B. immaculate and spotless

Step Five: Revise the Items

After constructing the initial pool of test items, the fifth step in the test construction process is to revise the items. This is best done by forming an item

review panel made up of qualified colleagues. Qualified colleagues would include those who have some general expertise in test construction and who are experts on the subject matter. The test constructor should ask the panel members to assess the items for accuracy, appropriateness, technical flaws, grammar, offensiveness, and level of readability (Crocker & Algina, 1986). It is generally recommended that about 20% more items be written than are needed so that after the revision process, an adequate number of items still exists for the final version of the test (Aiken, 1976).

Step Six: Pretest and Revise Items

Sixth, the items must be pretested and revised. The items should be taken by a small sample of examinees (usually 15 to 30 individuals) similar to those who will take the completed test. This pretest is not very formal; the test administrator should use this time to assess the examinee's reactions during testing, such as scribbling, long pauses, and answer-changing. After the examinees are finished, they should be invited to comment on each item and to offer suggestions for possible improvements. Item analysis can then be conducted to evaluate item difficulty (the percentage of examinees who correctly answer an item) and item discrimination power (the extent to which the item is answered correctly more often by those obtaining higher overall test scores than by those obtaining lower overall test scores). With these qualitative comments by examinees and quantitative results of item analysis, the best items can be chosen and revised if needed.

Lastly, the test constructor can assemble the final form of the test. It can be determined at this time that the constructor has a collection of good items. It will be determined if he/she has developed a good overall test after doing further analyses.

References

Aiken, L.R., Jr. (1976). Psychological testing and assessment (2nd ed.). Boston: Allyn and Bacon, Inc.

Bloom, B.S., Englehart, M.D., Furst, G.J., Hill, W.H., & Krathwohl, D.R. (1956). Taxonomy of educational objectives: Handbook I, The cognitive domain. New York: David McKay Co., Inc.

Brown, F.G. (1983). Principles of educational and psychological testing (3rd ed.). New York: Holt, Rinehart, and Winston.

Brown, F.G. (1976). Principles of educational and psychological testing (2nd ed.). New York: Holt, Rinehart and Winston.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Gay, L.R. (1980). Educational evaluation and measurement: Competencies for analysis and application. Columbus, OH: Charles E. Merrill Publishing Co.

Gronlund, N.E. (1977). Constructing achievement tests (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc.

Hopkins, K.D. & Stanley, J.C. (1981). Educational and psychological measurement and evaluation (6th ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc.